Apprentissage de structures séquentielles pour l'extraction d'entités et de relations dans des textes d'appels d'offres

Oussama Ahmia*, Nicolas Béchet*, Pierre-François Marteau*,

* IRISA, Université Bretagne Sud, Rue Yves mainguy – BP 573 56000 VANNES cedex
nom.prénom@irisa.fr, http://www-expression.irisa.fr

Résumé. Dans cet article nous présentons une étude exploitant des méthodes d'apprentissage automatique de structures séquentielles pour extraire des relations sémantiques dans des textes issus de bases d'appels d'offres. L'une des relations que nous considérons concerne l'emprise d'un projet d'aménagement, caractérisée par une association entre les concepts qui définissent les infrastructures (bâtiments) et les concepts qui définissent leur(s) surface(s) d'implantation. L'étude propose une analyse comparée d'approches à base de champs conditionnels aléatoires (CRF), de CRF d'ordre supérieur (H-CRF), de CRF semi-Markoviens, Modèles de Markov cachés (HMM) et de perceptrons structurés.

1 Introduction

L'identification des projets d'aménagement futurs permet de représenter la ville de demain, et pour cela, il est nécessaire de disposer en temps opportun de données sur la nature du projet d'aménagement, caractérisée en particulier par sa superficie et sa destination principale (typologie de bâtiment). Les appels d'offres relatifs aux marchés publics constituent ainsi une source importante d'informations. Sur la base d'un jeu de données collecté à partir du Bulletin Officiel des Annonces des Marchés Publics (BOAMP), notre étude concerne la détection de la typologie des bâtiments référencés dans une annonce et leurs superficies respectives.

L'extraction d'information que nous ciblons ici, consiste à extraire automatiquement des données structurées telles que des entités (nommées), des relations entre entités, ou encore des attributs décrivant des entités, à partir de sources non structurées (Sarawagi, 2008).

Dans le but d'extraire la surface d'un bâtiment dans des textes d'appels d'offres en français, nous proposons une analyse comparative d'un ensemble de modèles statistiques dont les Champs Aléatoires Conditionnels (CRF).

Différentes caractéristiques issues des textes sont exploitées : des caractéristiques grammaticales et sémantiques pour la caractérisations des entrées lexicales, et des ensembles de caractéristiques à *longue portée* (décrivant le contexte) qui ont démontré leur efficacité dans le domaine d'étiquetage des séquences symboliques (Li et al., 2011).